Cloudification of scientific experimental data using Onedata

Introduction

- In many scientific disciplines the expensive equipment is used. Data cloudification system should provide an easy way to make data **produced by a scientific experimental device** (e.g. cryoEM) **available to researchers** regardless of their institutions and geographical locality.
- Our solution uses the Onedata system as a backend data management system.
- It supports the whole process beginning from acquiring produced data from the device, its publishing according to required access policy and its archiving in permanent storage.

Onedata system

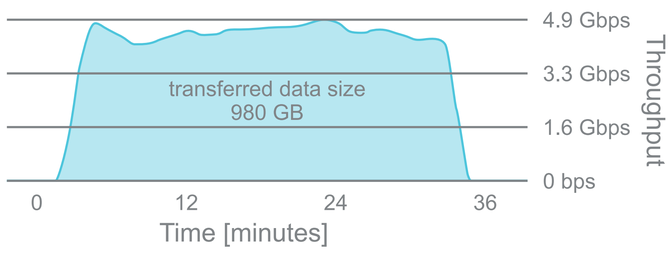
- Onedata is a high-performance data management solution that offers unified data access across globally distributed environments.
- Space - logical data container that integrates access to physical data located on different types of storage systems (POSIX, NFS, S3, Ceph, …) in different data providers, exposing a POSIX-like filesystem.
- Onedata supports data transfers and replication management, QoS rules, public data sharing, PID/DOI minting for OpenData purposes, metadata indexing and discovery, archiving for long-term preservation, and automated data processing using workflows.

EGI DataHub

datahub.egi.eu

- Controlled environment for data sharing
- Established and well-managed EGI service
- The **central Onezone instance** of the EGI Federation
- Single Sign On (SSO) through EGI Check-in
- Supported by EGI-ACE project

Transfer performance



Production use

- Currently over 1100 data Spaces
- In total 780 TB
- Running 4 Oneproviders
- Import data from 4 facilities
- Tens of users (and increasing)

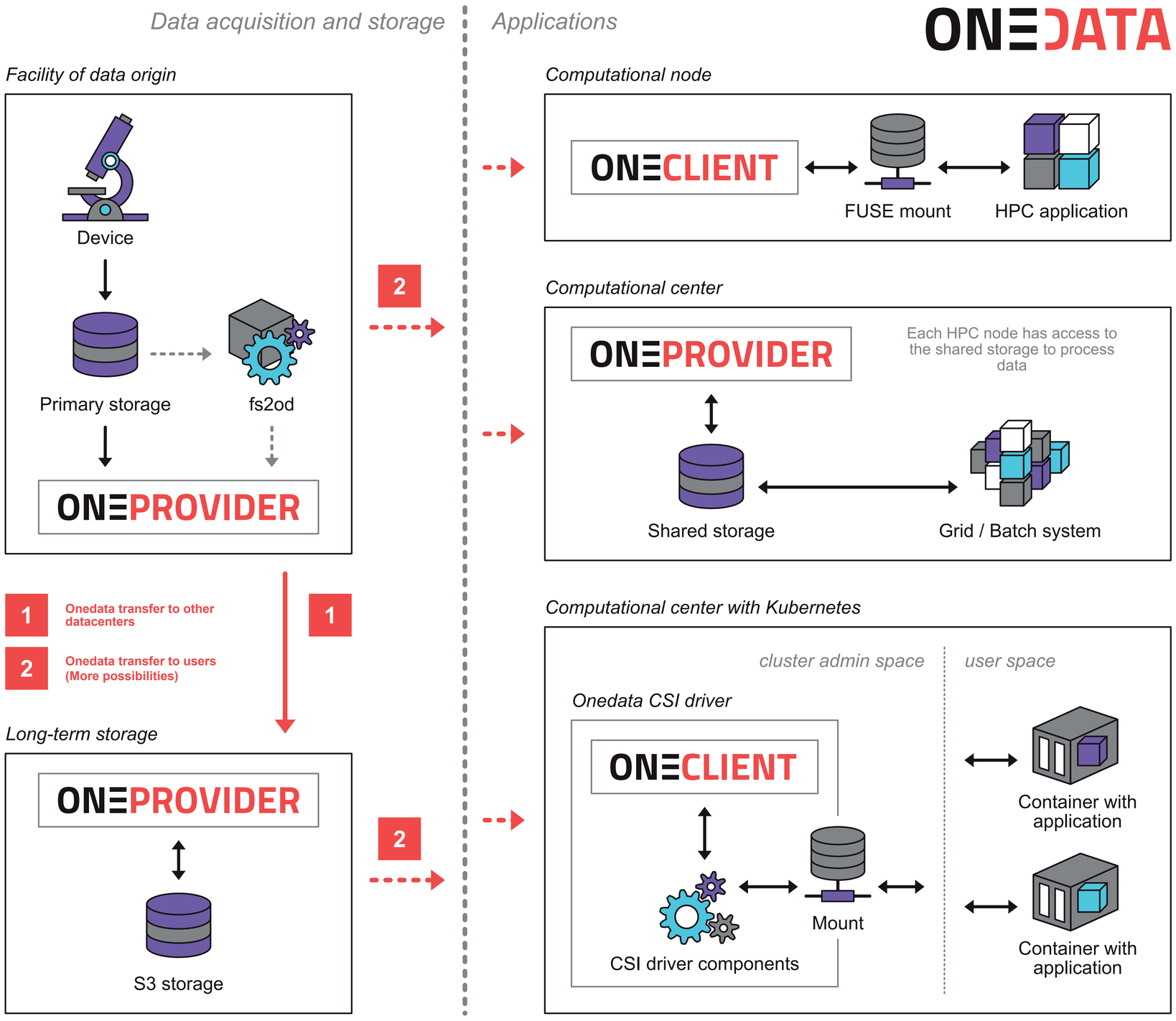
fs2od - filesystem to Onedata

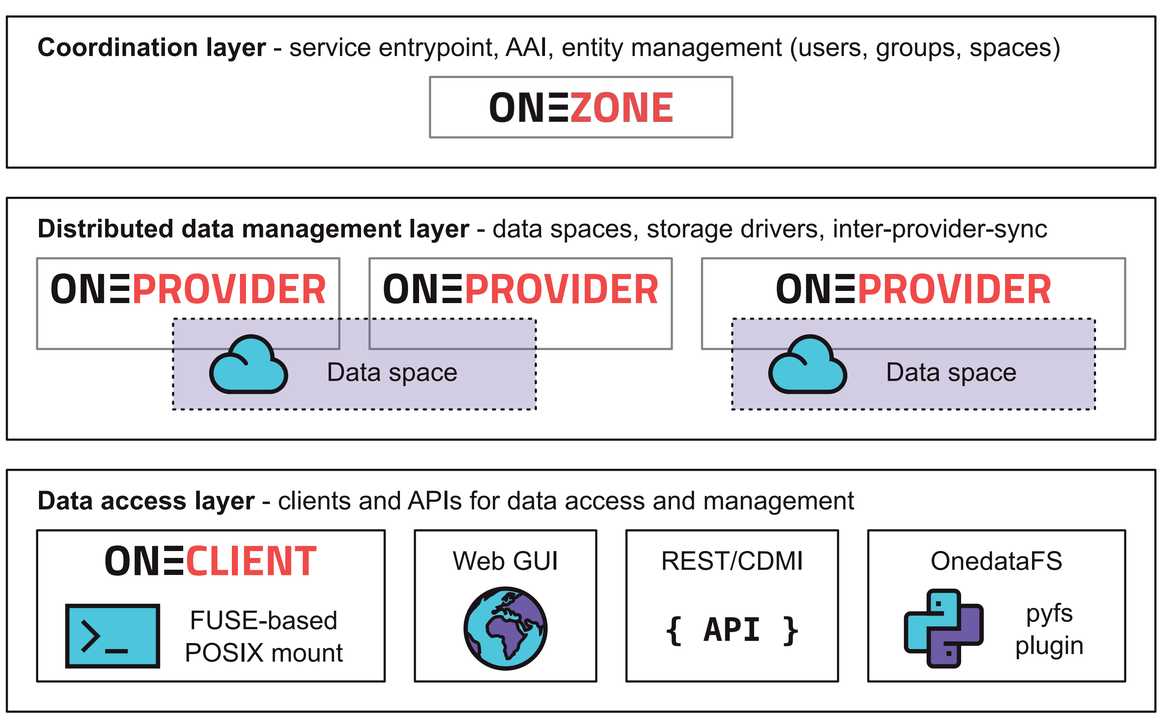
github.com/CERIT-SC/fs2od

- Containerized command-line utility
- Automatically register data from POSIX filesystem to Onedata
- Configurable options
- Metadata support & e-mail notifications
- Set up replication, archiving & removing policy

International deployment

- Data provider site (scientific facility)
- National data repository
- Community repositories (optional)
- Near-compute replicas





Onedata CSI driver

github.com/CERIT-SC/csi-onedata

- Implementation of the CSI driver (standard for exposing arbitrary block and file storage systems to containerized workloads) to **access Onedata in Kubernetes**.
- Use build-in components of Kubernetes for mounting miscellaneous types of storages instead of direct mount in PODs started by users.
- Oneclient requires capabilities, which are not suitable on shared Kubernetes clusters. Onedata CSI driver solves this problem. Driver has to be installed by cluster administrator, but user PODs don't need to be started with root or with hazardous capabilities. This solution is **friendly to Kubernetes security context**.

Onedata access information stored by dataset

- Automatically inserted to specified metadata file

```
onedata:
  onezone: https://datahub.egi.eu
  spaceId: c2956f8d21fffd7bcbb628b382f0c17bch2e97
  inviteToken: MDAxY2xvY2F00aW9uIGRhdGFodWIuZWdpLmV1CjAwOT...
  publicUrl: https://datahub.egi.eu/share/885c806b0c94730195fe65e...
```


Tomáš Svoboda[1,2], Aleš Křenek[1,2], Łukasz Opioła[3], Andrea Manzi[4], Josef Handl[1,2]

1. Masaryk University
2. CESNET
3. ACC Cyfronet AGH
4. EGI Foundation

svoboda@ics.muni.cz

cerit-sc.cz/onedata